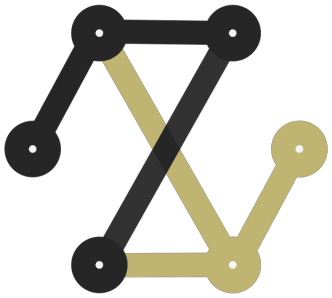

LibreIM

Introducción a PAC Learning

Ignacio Cordón Castillo

8 de enero de 2017

Índice

Estos apuntes son una adptación en su mayoría del contenido del libro [1]

1. Introducción

Damos unas notaciones/definiciones básicas que utilizaremos de aquí en adelante.

- **Dominio**: $\mathcal{X}$. LLamamos una instancia a $x \in \mathcal{X}$
- **Conjunto de etiquetas**: $\mathcal{Y}$ consideramos $\{0, 1\}$, lo que nos restringe al paradigma binario.
- **Verdadero etiquetado**: Asumimos la existencia de una función $f : \mathcal{X} \to \mathcal{Y}$ que devuelve el verdadero etiquetado de todas las instancias.
- **Generación de instancias**: Asumimos la existencia de una distribuciónd de probabilidad $\mathcal{D}$ sobre $\mathcal{X}$ que nos da información sobre la probabilidad de extraer cada posible instancia desde $\mathcal{X}$.
- **Conjunto/Secuencia de entrenamiento**: $S = ((x_1, y_1), \dots (x_m, y_m))$ secuencia con cada elemento perteneciente a $\mathcal{X} \times \mathcal{Y}$. A veces lo llamaremos conjunto, por abuso de notación, pero se trata de una secuencia en la que pueden repetirse tuplas. Los ejemplos del conjunto de entrenamiento representan una m.a.s $(\mathcal{X}_1, \dots \mathcal{X}_m)$, muestra aleatoria simple, idénticamente distribuida, donde cada X_i sigue la misma distribución que $\mathcal{X}$, $X_i \sim \mathcal{D}$. Además, cada ejemplo del conjunto de entrenamiento se etiqueta según f. Notamos este hecho $S \sim \mathcal{D}^m$.
- **Resultado del aprendizaje**: una función $h : \mathcal{X} \to \mathcal{Y}$ que llamaremos hipótesis/clasificador. Se usa la notación $A(S)$ para denotar la hipótesis que un algoritmo A devuelve para una secuencia de entrenamiento S.
- **Error del clasificador**: Sea $A \subset \mathcal{X}$ miembro de alguna σ -álgebra de conjuntos de $\mathcal{X}$, $\mathcal{D}$ distribución de probabilidad sobre $\mathcal{X}$, $\mathcal{D}(A)$ denota la probabilidad de que un punto de $\mathcal{X}$ esté en A, es decir, la probabilidad de que ocurra el hecho A. Basándonos en este, definimos el error del clasificador h como:

$$L_{D,f}(h) := \mathcal{D}(\{h(x) \neq f(x)\}) = \mathbb{P}_{x \sim \mathcal{D}}[h(x) \neq f(x)]$$

Además asumimos que el algoritmo de aprendizaje no conoce ni la distribución $\mathcal{D}$ ni la función f.

1. Minimización del riesgo empírico (ERM)

   Definición 1. ***Riesgo empírico (ER)***

   Definimos el riesgo empírico o error empírico como:

   $$L_S(h) = \frac{|i \in 1 \dots m : h(x_i) \neq y_i|}{m}$$

   Podemos pensar en él como el error del clasificador sobre el conjunto de entrenamiento. El paradigma que intenta buscar una hipótesis que minimice el error empírico recibe el nombre de *Minimización de Riesgo Empírico - ERM* y notamos $ERM(S)$ al clasificador que obtenemos basándonos en este paradigma para un determinado conjunto de entrenamiento S.

   Este error no es siempre óptimo. Pensemos en el siguiente ejemplo:

   Sea $\mathcal{X} = \mathbb{R}$, $\mathcal{D}$ la distribución uniforme sobre $[0, 2] \subset \mathbb{R}$, y la siguiente función:

   $$f(x) = \begin{cases} 1 & x \in [0, 1] \\ 0 & x \in \mathbb{R} \setminus [0, 1] \end{cases}$$

   $S = \{(x_1, y_1), \dots (x_m, y_m)\}$ un conjunto de entrenamiento de tamaño m sin elementos repetidos y el clasificador:

   $$h_S(x) = \begin{cases} y_i & \exists i \in \{1 \dots m\} : x = x_i \\ 0 & \nexists i \in \{1 \dots m\} : x = x_i \end{cases}$$

   Este clasificador es perfecto respecto a la minimización de riesgo empírico, pero $\mathbb{P}_{x \sim \mathcal{D}}[h_S(x)] = 1/2$, es decir, tiene el mismo nivel de acierto que el clasificador idénticamente 1. A este fenómeno lo denominamos **overfitting**.

2. ERM con *sesgo inductivo*

   Se intenta corregir el ERM corrigiendo el espacio de búsqueda, esto es, la clase de hipótesis $\mathcal{H}$ desde la que el algoritmo puede escoger un $h : \mathcal{X} \to \mathcal{Y}$. Llamamos a esto *sesgo inductivo* puesto que se asumirá una determinada clase de funciones $\mathcal{H}$ en función de las características del problema.

   Notaremos a este nuevo paradigma $ERM_{\mathcal{H}}(S)$, y lo definimos de manera que:

$$ERM_{\mathcal{H}}(S) := h_S \in argmin_{h \in \mathcal{H}} L_S(h)$$

Definimos la propiedad de factibilidad, que usaremos más adelante.

Definición 2. ***Propiedad de factibilidad***

Existe $\bar{h} \in \mathcal{H}$ *verificando* $L_{D,f}(\bar{h}) = 0$.

La hipótesis de factibilidad implica que $\mathbb{P}_{S \sim \mathcal{D}^m}[L_S(\bar{h}) = 0] = 1$, y por tanto $\mathbb{P}_{S \sim \mathcal{D}^m}[L_S(h_S) = 0] = 1$.

El valor $L_{\mathcal{D},f}(h_S)$ dependerá del conjunto de entrenamiento S, y la elección del mismo está sometida al azar. Además, necesitamos definir cómo de buena será la predicción.

2. Aprendizaje PAC.

Definición 3. ***Aprendizaje PAC (Probablemente Aproximadamente Correcto)***

Una clase de funciones $\mathcal{H}$ es PAC learnable si existe una función $m_{\mathcal{H}} :]0,1[^2 \to \mathbb{N}$, llamada complejidad muestral, y un algoritmo A verificando que si $0 \leq \epsilon, \delta \leq 1$, entonces para toda distribución $\mathcal{D}$ sobre $\mathcal{X}$ y para toda función de verdadero etiquetado $f : \mathcal{X} \to \{0,1\}$, si la propiedad de factibilidad se cumple, ejecutando el algoritmo para un conjunto de entrenamiento $S \sim \mathcal{D}^m$ etiquetado mediante f, con $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ el algoritmo devuelve una hipótesis $A(S) = h \in \mathcal{H}$ verificando que:

$$\mathbb{P}_{S \sim \mathcal{D}^m}[L_{\mathcal{D},f}(h) \leq \epsilon] \geq 1 - \delta$$

$(1 - \delta)$ es la *confianza de la predicción* (probablemente) y ϵ la *exactitud* (correcto).

Podemos considerar $m_{\mathcal{H}}$ única en el sentido de que para cada (δ, ϵ) nos devuelve el menor natural verificando las hipótesis del enunciado.

Teorema 1. ***Las clases finitas de funciones son PAC learnable***

Sea $\mathcal{H}$ una clase de funciones finita. Sean $0 \leq \epsilon, \delta \leq 1$, y un natural $m \in \mathbb{N}$ verificando:

$$m \geq \frac{1}{\epsilon} log\left(\frac{|\mathcal{H}|}{\delta}\right)$$

Entonces para toda función de verdadero etiquetado $f : \mathcal{X} \to \{0,1\}$, y para toda distribución $\mathcal{X} \sim \mathcal{D}$ para la que se verifique la ***propiedad de factibilidad*** *entonces las hipótesis que obtenemos a través del algoritmo ERM son con una confianza superior a $1 - \delta$ ϵ exactas.*

Como consecuencia, deducimos que la complejidad muestral es menor o igual a $\left\lceil \frac{1}{\epsilon} log\left(\frac{|\mathcal{H}|}{\delta}\right) \right\rceil$

Demostración. Fijada una distribución $\mathcal{D}$ y una función de etiquetado f, notamos:

$$\mathcal{H}_B = \{h \in \mathcal{H} : L_{\mathcal{D},f}(h) > \epsilon\}$$

Se tiene:

$$\mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D},f}(h_S) > \epsilon] \leq \mathbb{P}_{S\sim\mathcal{D}^m}[\exists h \in \mathcal{H}_B : L_S(h) = 0] \leq \sum_{h\in\mathcal{H}_B} \mathbb{P}_{S\sim\mathcal{D}^m}[L_S(h) = 0]$$

La primera desigualdad viene dada porque dada h_S se verifica, por la propiedad de factibilidad, que $L_S(h_S) = 0$. La segunda por subaditividad.

Además, fijada $h \in \mathcal{H}_B$, como $L_{\mathcal{D},f}(h) > \epsilon$:

$$\mathbb{P}_{S\sim\mathcal{D}^m}[L_S(h) = 0] = \mathbb{P}_{(x_1,\dots x_n)\sim\mathcal{D}^m}[\forall i \quad h(x_i) = f(x_i)] =$$
$$= \prod_{i=1}^{m} \mathbb{P}_{x\sim\mathcal{D}}[h(x) = f(x)] = \prod_{i=1}^{m}(1 - L_{\mathcal{D},f}(h)) \leq (1-\epsilon)^m \leq e^{-\epsilon m}$$

Las dos desigualdades probadas, junto a la hipótesis del enunciado, y usando $\mathcal{H}_B \subseteq \mathcal{H}$ dan lugar a:

$$\mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D},f}(h_S) > \epsilon] \leq |\mathcal{H}|e^{-\epsilon m} \leq \delta$$

□

3. Generalización aprendizaje PAC: PAC agnóstico

Hasta ahora tenemos dos problemas en la definición de PAC. Intentamos buscar una hipótesis sobre una función de verdadero etiquetado, f, que por tanto no podrá asignar dos imágenes distintas al mismo punto, y además, estamos suponiendo que se cumple la propiedad de factibilidad.

Para paliar esto, podríamos considerar $\mathcal{D}$ como la distribución conjunta sobre $\mathcal{X} \times \mathcal{Y}$, y la noción de error para $h : \mathcal{X} \to \mathcal{Y}$ quedaría:

$$L_{\mathcal{D}}(h) := \mathbb{P}_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$$

Con estos conceptos revisitados, podríamos asegurar que la hipótesis que menor error comete para $\mathcal{Y} = \{0, 1\}$ es el llamado **clasificador de Bayes**:

$$f_{\mathcal{D}}(x) = \begin{cases} 1 & \mathbb{P}[y = 1|x] >= 0{,}5 \\ 0 & si \quad no \end{cases}$$

Pero deseamos ir aún más allá, y poder generalizar la definición para una función de pérdida arbitraria.

Definición 4. ***Función de pérdida***

Dados un conjunto $\mathcal{H}$, Z y una σ álgebra de conjuntos sobre Z, se denomina función de pérdida de $\mathcal{H}$ sobre Z a cualquier función de la forma:

$$l : \mathcal{H} \times Z \to \mathbb{R}^+$$

que verifique que la función currificada $l(h, \cdot)$ sea medible $\forall h \in \mathcal{H}$ sobre la σ álgebra inicial.

Con funciones de pérdidas arbitrarias, redefiniríamos los conceptos de *error* y *error empírico* de la forma:

$$L_{\mathcal{D}}(h) := \mathbb{E}_{z\sim\mathcal{D}}[l(h,z)]$$
$$L_S(h) := \frac{1}{m}\sum_{i=1}^{m} l(h,z_i)$$

Definición 5. ***Aprendizaje PAC agnóstico***

Una clase de funciones $\mathcal{H}$ es agnósticamente PAC learnable respecto a Z (sobre el que tenemos definida una σ álgebra de conjuntos) y a una función de pérdida $l : \mathcal{H} \times Z \to \mathbb{R}^+$ si existe una función $m_{\mathcal{H}} :]0,1[^2 \to \mathbb{N}$ y un algoritmo A verificando que si $0 \leq \epsilon, \delta \leq 1$, entonces para toda distribución $\mathcal{D}$ sobre Z ejecutando el algoritmo para un conjunto de entrenamiento $S \sim \mathcal{D}^m$, con $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ el algoritmo devuelve una hipótesis $A(S) = h \in \mathcal{H}$ verificando que:

$$\mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(h) \leq inf_{h'\in\mathcal{H}} L_{\mathcal{D}}(h') + \epsilon] \geq 1 - \delta$$

Notamos que esta definición, en caso de cumplirse la propiedad de factibilidad, tomando $Z = \mathcal{X} \times \mathcal{Y}$, y la llamada función de pérdida 0-1:

$$l_{0-1}(h(x,y)) := \begin{cases} 0 & h(x) = y \\ 1 & si \quad no \end{cases}$$

equivale a la primera definición que dimos de aprendizaje PAC. Por ello no distinguiremos en el uso de uno u otro concepto, sino que se deducirá de si estamos asumiendo propiedad de factibilidad o no.

Cuando permitimos que el algoritmo A devuelva una función $h \notin \mathcal{H}$, de manera que $h \in \mathcal{H}'$ y $\mathcal{H} \subset \mathcal{H}'$ una clase de funciones a donde la función de pérdida es extendible de manera natural, el aprendizaje recibe el nombre de **aprendizaje impropio**. La definición aquí dada se ha hecho para **aprendizaje propio**.

4. Condiciones suficientes para ser PAC learnable

Definición 6. ***Conjunto de entrenamiento ϵ representativo***

Un conjunto de entrenamiento S se dice ϵ representativo respecto a un dominio Z, a una clase de hipótesis $\mathcal{H}$, una función de pérdida l y una distribución $\mathcal{D}$ si:

$$\forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon$$

Lema 1. *Sea un conjunto de entrenamiento de tamaño S, $\frac{\epsilon}{2}$ representativo respecto a un dominio Z, a una clase de hipótesis $\mathcal{H}$, una función de pérdida l y una distribución $\mathcal{D}$. Entonces:*

$$L_{\mathcal{D}}(ERM(S)) \leq inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$$

Demostración. Para $h \in \mathcal{H}$ arbitrario.

$$L_{\mathcal{D}}(h_S) \leq L_{\mathcal{S}}(h_S) + \frac{\epsilon}{2} \leq L_{\mathcal{S}}(h) + \frac{\epsilon}{2} \leq L_{\mathcal{D}}(h) + \frac{\epsilon}{2} + \frac{\epsilon}{2} = L_{\mathcal{D}}(h) + \epsilon$$

□

Definición 7. ***Convergencia uniforme***

Decimos que una clase de hipótesis $\mathcal{H}$ tiene la propiedad de convergencia uniforme respecto a un dominio Z, y a una función l si para todo $0 < \delta, \epsilon < 1$ existe $m_{\epsilon,\delta}$ verificando que para toda distribución $\mathcal{D}$ sobre Z, si S es un conjunto de entrenamiento de tamaño mayor o igual a $\$m_{\epsilon,\ \delta}$, entonces:

$$P_{S \sim \mathcal{D}^m}[\forall h \in \mathcal{H} | L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon] \geq 1 - \delta$$

Teorema 2. ***La convergencia uniforme es condición suficiente para ser PAC learnable***

Sea $\mathcal{H}$ una clase de hipótesis con la propiedad de convergencia uniforme. Entonces es PAC learnable con complejidad muestral menor o igual al $m_{\frac{\epsilon}{2},\delta}$ dado en la definición anterior y el algoritmo ERM

Proposición 1. ***Las clases finitas tienen la propiedad de convergencia uniforme***

Sea $\mathcal{H}$ una clase de hipótesis finita, Z un dominio y sea $l : \mathcal{H} \times Z \to [a, b]$ una función de pérdida. Entonces $\mathcal{H}$ verifica la propiedad de convegencia uniforme con:

$$m_{\epsilon,\delta} \leq \left\lceil \frac{log(2|\mathcal{H}|/\delta)(b-a)^2}{2\epsilon^2} \right\rceil$$

Lema 2. ***Desigualdad de Hoeffding***

Sean $X_1, \dots X_n$ una muestra aleatoria simple de una variable X, $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$ con $E[\bar{X}] = \mu$ y $P[a \leq X_i \leq b] = 1$. Entonces para todo $\epsilon > 0$

$$P\left[\left|\bar{X} - \mu\right| > \epsilon\right] \leq 2e^{-2m\left(\frac{\epsilon}{(b-a)}\right)^2}$$

Demostración. Sea $\mathcal{H}$ una clase de hipótesis finita.

Fijamos $0 < \delta, \epsilon < 1$. Necesitamos encontrar $m \in \mathbb{N}$ verificando:

$$P_{S\sim\mathcal{D}^m}[\exists h \in \mathcal{H} | L_S(h) - L_{\mathcal{D}}(h)| > \epsilon] < \delta$$

Partimos de la siguiente desigualdad, que usaremos más adelante, obtenida por subaditividad:

$$P_{S\sim\mathcal{D}^m}[\exists h \in \mathcal{H} | L_S(h) - L_{\mathcal{D}}(h)| > \epsilon] \leq \sum_{h\in\mathcal{H}} P_{S\sim\mathcal{D}^m}[|L_S(h) - L_{\mathcal{D}}(h)| > \epsilon]$$

Fijamos $h \in \mathcal{H}$.

Dado un conjunto de entrenamiento $S = (z_1, \dots z_m)$, recordamos que $L_{\mathcal{D}}(h) = \mathbb{E}_{z\sim\mathcal{D}}[l(h, z)]$ y que $L_S(h) = \frac{1}{m}\sum_{i=1}^{m} l(h, z_i)$

Donde $z_i \sim \mathcal{D}$ y por tanto $\mathbb{E}_{S\sim\mathcal{D}^m}[L_S(h)] = \mathbb{E}_{z\sim\mathcal{D}}[l(h, z)] = L_{\mathcal{D}}(h)$. Además, llamando $X_i = l(h, z_i)$, por ser z_i realizaciones muestrales de una m.a.s

se tiene que las X_i son independientes e idénticamente distribuidas, con $P[a < X_i < b] = 1$. Estamos en condiciones de aplicar la desigualdad de Hoeffding.

Por tanto:

$$P_{S\sim\mathcal{D}^m}\left[\left|\frac{1}{m}\sum_{i=1}^{m}X_i - L_{\mathcal{D}}(h)\right| > \epsilon\right] = P_{S\sim\mathcal{D}^m}[|L_S(h) - L_{\mathcal{D}}(h)| > \epsilon] \leq 2e^{-2m\left(\frac{\epsilon}{b-a}\right)^2}$$

Y por tanto:

$$P_{S\sim\mathcal{D}^m}[\exists h \in \mathcal{H} | L_S(h) - L_{\mathcal{D}}(h)| > \epsilon] < |\mathcal{H}|2e^{-2m\left(\frac{\epsilon}{b-a}\right)^2}$$

□

Recordemos hasta ahora el resultado que habíamos obtenido era su carácter PAC learnable, donde agnósticamente PAC learnable y learnable con funciones de pérdida 0-1 era un término equivalente. El teorema que enunciamos a continuación, deducible a partir del teorema sobre el caracter agnóstico - PAC learnable de clases de funciones con propiedad de convergencia uniforme, en particular las finitas, generaliza el resultado para cualquier funciones de pérdida acotada.

Teorema 3. ***Las clases finitas son agnósticamente PAC learnable***

Sea $\mathcal{H}$ una clase de hipótesis finita, Z un dominio y sea $l : \mathcal{H} \times Z \to [a, b]$ una función de pérdida. Entonces $\mathcal{H}$ es PAC learnable con complejidad muestral:

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{2log(2|\mathcal{H}|/\delta)(b-a)^2}{\epsilon^2} \right\rceil$$

5. Equilibrio error-varianza *bias-complexity tradeoff*

Veamos que dado un algoritmo de aprendizaje no puede ser el óptimo para aprender todas las distribuciones.

Damos un lema previo, la desigualdad de Markov:

Lema 3. ***Desigualdad de Markov***

Dada una variable aleatoria Z no negativa. Entonces para todo $a \geq 0$

$$P[Z \geq a] \leq \frac{\mathbb{E}[Z]}{a}$$

Teorema 4. ***Teorema de No Free Lunch***

Sea A cualquier algoritmo de aprendizaje para clasificación binaria con respecto a la función de pérdida 0-1 sobre el dominio $\mathcal{X}$. Sea un conjunto de entrenamiento de tamaño $m < |\mathcal{X}|/2$. Entonces existe una distribución $\mathcal{D}$ sobre $\mathcal{X} \times \{0,1\}$ verificando:

1. *Existe una función $f : \mathcal{X} \to \{0,1\}$ con $L_{\mathcal{D}}(f) = 0$*
2. *$P_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) \geq 1/8] \geq 1/7$*

Demostración. Sea un conjunto de entrenamiento (consideramos un conjunto y no una secuencia) de tamaño $2m$, C. Hay $T = 2^{2m}$ posibilidades de etiquetado del conjunto, esto es, 2^{2m} posibles hipótesis, $f_i : C \to \{0,1\}$, que vamos a extender a $\mathcal{X}$ llamándolas $\bar{f}_i$ de forma que $\bar{f}_{i|C} = f_i$ y $\bar{f}_i(x) = 0 \quad \forall x \in \mathcal{X} \setminus C$. Vamos a tomar para cada una de ellas una distribución $\mathcal{D}_i$ definida sobre $\mathcal{X} \times \{0,1\}$ definida por:

$$\forall (x,y) \in \mathcal{X} \times \{0,1\} \qquad P_{z\sim\mathcal{D}_i}[z = (x,y)] = \begin{cases} 1/|C| & \exists x_i \in C : y = f(x_i) \\ 0 & si \quad no \end{cases}$$

Claramente $L_{\mathcal{D}_i}(f_i) = 0$

Vamos a probar que:

$$\exists i \in \{1, \dots 2m\} : \mathbb{E}_{S \sim \mathcal{D}_i^m}[L_{\mathcal{D}_i}(A(S))] \geq \frac{1}{4}$$

Hay $k = (2m)^m$ posibles secuencias de entrenamiento de tamaño m, $S_j, j = 1, \dots k$ tomadas desde C. Siendo $S_j = (x_1, \dots x_m)$ notamos $S_j^i = ((x_1, f_i(x_1)), \dots, (x_m, f_i(x_m)))$. Cada S_j tiene la misma probabilidad de ser nuestro conjunto de entrenamiento (extracción de m valores con reemplazamiento desde el conjunto C), verificándose:

$$\mathbb{E}_{S \sim \mathcal{D}_i^m}[L_{\mathcal{D}_i}(A(S))] = \frac{1}{k} \sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i))$$

Recordando que hemos llamado $k = (2m)^m$, $T = 2^{2m}$, se tiene:

$$\begin{aligned} max_{i \in \{1, \dots T\}} \frac{1}{k} \sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i)) &\geq \frac{1}{T} \sum_{i=1}^{T} \frac{1}{k} \sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i)) = \\ &= \frac{1}{k} \sum_{j=1}^{k} \frac{1}{T} \sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i)) \geq \\ &\geq min_{j \in \{1, \dots k\}} \frac{1}{T} \sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i)) \end{aligned}$$

Además fijado $j \in \{1, \dots k\}$, se tiene que que para todo $i \in \{1, \dots T\}$:

$$L_{\mathcal{D}_i}(h) = \frac{1}{|C|} \sum_{x \in C} \mathbb{1}_{[A(S_j^i)(x) \neq f_i(x)]} = \frac{1}{2m} \sum_{x \in C} \mathbb{1}_{[A(S_j^i)(x) \neq f_i(x)]}$$

Por tanto:

$$\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i)) \geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{2m}\sum_{x\in C}\mathbb{1}_{[A(S_j^i)(x)\neq f_i(x)]} =$$

$$= \frac{1}{2m}\sum_{x\in C}\frac{1}{T}\sum_{i=1}^{T}\mathbb{1}_{[A(S_j^i)(x)\neq f_i(x)]} \geq$$

$$\geq \frac{1}{2} min_{x\in C}\frac{1}{T}\sum_{i=1}^{T}\mathbb{1}_{[A(S_j^i)(x)\neq f_i(x)]}$$

Como dado un $x \in C$ cualquiera, la mitad de clasificadores f_i clasificarán x bien y la otra mitad mal, se tiene:

$$\frac{1}{2} min_{x\in C}\frac{1}{T}\sum_{i=1}^{T}\mathbb{1}_{[A(S_j^i)(x)\neq f_i(x)]} = \frac{1}{2}\frac{1}{T}\frac{T}{2} = \frac{1}{4}$$

Y uniendo toda esta información:

$$max_{i\in\{1,\dots T\}}\frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i)) \geq \frac{1}{4}$$

Sea $k = argmax_{i\in\{1,\dots T\}}\frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i))$

Si $\mathcal{D} = \mathcal{D}_k$ cumple la parte 2 del enunciado del teorema, es nuestra distribución buscada, y como función buscada en el apartado 1. podemos tomar $f = f_k$

Como $L_{\mathcal{D}}(A(\cdot))$ puede ser vista como una variable aleatoria donde $S \sim \mathcal{D}^m$ y que toma valores en $[0, 1]$, tenemos que tomando $Z = 1 - L_{\mathcal{D}}(A(\cdot))$, $a = \frac{7}{8}$ en el lema previo llegamos a:

$$P_{S\sim\mathcal{D}^m}\left(\frac{1}{8} \geq L_{\mathcal{D}}(A(S))\right) \leq \frac{3}{4}\cdot\frac{8}{7} = 24/28$$

donde $\mathbb{E}(Z) = \mathbb{E}(1 - L_{\mathcal{D}}(A(\cdot))) = 1 - \mathbb{E}(L_{\mathcal{D}}(A(\cdot))) \leq \frac{3}{4}$

Es decir:

$$P_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}(A(S)) \geq \frac{1}{8}\right) \geq \frac{4}{28} = \frac{1}{7}$$

□

Como consecuencia del teorema, podemos decir que no hay un algoritmo de aprendizaje óptimo para todas las distribuciones, puesto que para una dada por el resultado del teorema, el algoritmo ERM con $\mathcal{H} = \{f\}$ aprendería mejor.

Referencias

[1] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning*, ser. nil. Cambridge University Press, 2014. [Online]. Available: http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning